Service Oriented Speech Control Robot System

ZHU Qingbo[1], Liu Yaohe[2], Song Tingxin[3]

Department of Mechanical Engineering, Hubei University of Technology

[1]zhuqingbo@msn.com; [2]yaohe_liu@126.com; [3]songtx2006@163.com

***Abstract*—A distributed remote speech control system based on Web Service was introduced. The system enables the transition of an inbound call via telephone into robot commands by the speech recognition engine and TTS speech synthesis engine and with the Web Service components, then the system transmits the real-time command into a robot and after receiving and recognizing the command, the robot moves its arm accordingly so as to achieve the goal of controlling the robot movement. The innovation with this system lies in the sufficient application of telephone and computer network to realize a distributed speech control.**

***Keywords*—Speech Recognition; Robot; Web Service; TTS; gSOAP**

I. INTRODUCTION

With the development of speech recognition technology and the internet, speech control and remote control have been attached more and more importance. A speech control system based on Web Service has become a burgeoning hotspot for robot control. However, how to integrate large varieties of robots in manufacturing industry and household appliance through internet with speech control technology has always been an outstanding question in automation industry. As a brand-new network control pattern, remote speech control system based on Web Service, from its inception, has exhibited a powerful vitality. Its features of a consummate integration of telephone and computer network, inclusion of several distributed application have won favors from many technical experts and users.

Robot speech recognition, namely the robot's hearing is that the robot receives a human signal, identifying its implication and adjusts its movement accordingly. The robot's hearing symbolizes the intelligence of the robot. Web Service based remote control is defined as the extension of a human sense organ to a long distance robot through internet to accomplish a task. This research study has adopted CTI (Computer Telephone Intercommunication) speech recognition technology to realize a remote control over an industry robot through a variety of logic movements based on Web Service components. Based on distributed network architecture, this system has revolutionized the traditional way of P2P control. This not only provides a mutual interactive setting for remote control, but also inaugurates a fresh industry for internet application. [1, 2, 3, 7]

II. SYSTEM ARCHITECTURE

This remote control system based on speech recognition technology has extended the local speech of the robot to a far-terminal so that the operator is able to manipulate the robot in a direct manner either by using modern computer communication technology or the telephone network without any sense of restrictions as shown in figure 1.

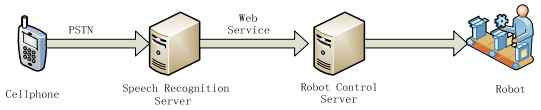

Figure 1. System flow chart

In the system architecture, we can see that the system has integrated the telephone and computer network so that the user can pass his speech command onto the recognition terminal through the telephone network, then the recognition-terminal will send the command into the control end of the robot through the computer network, as shown in figure 2. As it shown from the chart below, this entire system is consists of two parts, speech recognition server and robot control server. The speech recognition server works to receive speech signal from user's mobile and recognize the meaning of specify command form user which will be fully illustrate in section III. The remote robot control server works to receive result from speech recognition server and control the mechanical part of GT robot which is based on Web Service technology.

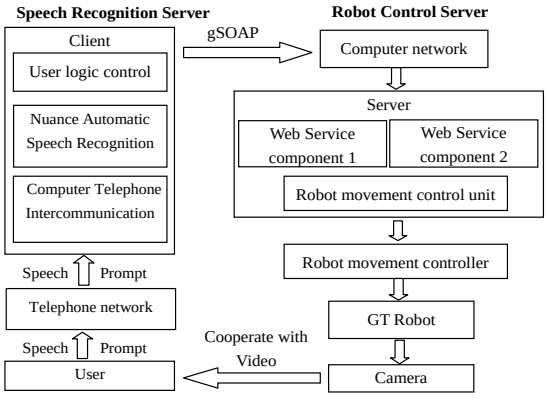

Figure 2. System flow chart

The remote mainframe is based on Web Service technology. On the one hand, as an agent for the interaction between the operator and the robot, Web Service components' function is a bridge to receive the speech commands from the operator while collect the real-time information from the robot. On the other hand, the remote mainframe, as an individual entity, namely an independent intelligent entity which acts on itself and has a sense of environment, is able to adjust according to the present environment so as to accomplish the tasks commanded by the operator in a long distance. In the speech control end, the computer network works as a medium for the interaction between the operator and the robot. The operator away from the sites is able to complete a task through the speech control over the robot and he can collect the parameters regarding the work conditions of the robot from the Web Service components as well.

III. SPEECH RECOGNITION MODEL

Figure1. HMM (Hidden Markov Model)

HMM, namely Hidden Markov Model, is a statistic model for the time sequences of speech recognition which can be seen as a dual random process in math: in one aspect, It can be seen as a hidden random process simulating the changes of speech signal by using Markov chain of limited status figure, in another it is an observation sequence relevant to every single Markov chain. The latter embodies the former while the former parameters are unpredictable. The human speech is actually a dual random process, and the speech signal is a predictable time sequence, which is generated as a parameter flow by the brain according to grammar and speech rules. Therefore HMM is an appropriate simulation of this process which has well displayed the entire instability and partial stability of speech signals. On the whole, human speech is an instable random process, yet when divided into several parts, it is available for linear analysis on a short-time basis [2, 3].

If construct a HMM based on this speech signals, we are able to identify different transient stable signal splits and track down their transition so as to complete the model construction based on the speech velocity and acoustic changes.

A speech recognition system or human speech recognition system based on HMM: We wish to calculate the probability of the observation sequence, $O = O_1 O_2 O_3 \cdots O_T$,given the model λ ,i.e., $P(O|\lambda)$ .The most straightforward way of doing this is through enumerating every possible state sequence of length T(the number of observations).Consider one such fixed state sequence

$$Q = q_1 q_2 \cdots q_T$$

Where q_1 is the initial state? The probability of the observation sequence O for the state sequence of (1) is

$$p(O \mid Q, \lambda) = \prod_{t=1}^{T} P(O_t \mid q_t, \lambda)$$

where we have assumed statistical independence of observations. Thus we get

$$p(O \mid Q, \lambda) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T)$$

The probability of such a state sequence Q can be written as

$$P(Q \mid \lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$$

The joint probability of O and Q , i.e., the probability that O and Q occur simultaneously, is simply the product of the above two terms, i.e.

$$P(Q \mid \lambda) = P(O, Q \mid \lambda) \bullet P(Q \mid \lambda)$$

The probability of O (given the model) is obtained by summing this joint probability over all possible state sequences q giving

$$P(O \mid \lambda) = \sum_{\forall Q} P(O \mid Q, \lambda) \bullet P(Q \mid \lambda) =$$

$$= \sum_{q_1 q_2 q_3 \cdots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{qT}(O_T)$$

The interpretation of the computation in the above equation is the following. Initially (at time t=1) we are in state q_1 with probability π_{q_1} , and generate the symbol O_1 (in this state) with probability $b_{q_1}(O_1)$. The clock changes from time t to t+1(t=2) and we make a transition to state q_2 from state q_1 with probability $a_{q_1 q_2}$, and generate symbol Q_2 with probability $b_{q_2}(O_2)$ .This process continues in this manner until we make the list transition (at time T) from state q_{T-1} to state q_T with probability $a_{q_{T-1} q_T}$ and generate symbol O_T with probability $b_{qT}(O_T)$.

Apparent, it is useless and inefficient to calculate the probability of the observation sequence. In this system, the recognition model has already taken some amends to this arithmetic and take some compensate to the noise exist in the environment [4, 5].

Figure2. Speech recognition engine

A general model for speech recognition system for several people is shown in figure 3. The model consists of four blocks. The first is data extraction that converts a wave data stored in audio wave format into a form that is suitable for further computer processing and analysis. The second is pre-processing, which involves filtering, removing pauses, silences and weak unvoiced sound signal and detect the valid speech signal. The third block is feature extraction, where speech features are extracted from the speech signal. The selected features have enough information to recognize a speaker. Here a class label is assigned to each word uttered by each speaker by examining the extracted features and comparing them with classes learnt during the training phase. Vector quantization is used as an identifier.

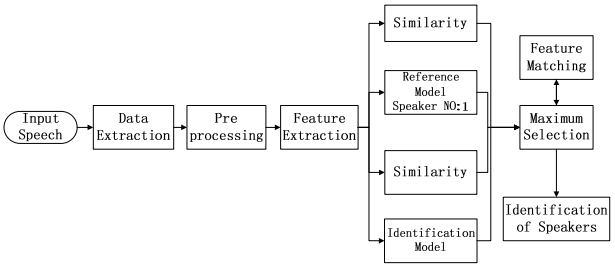

Figure 3. Speech signal process flow.

Speech recognition engine has provided two kinds of working patterns, namely recognition pattern and the command pattern. Different working pattern has decisive t impact on the recognition program set in it.

In this system, we have adopted the command pattern. In other words, we give the system the command to recognize the speech signal, and based on the speech, application programs and the recognition environment settings are designed. As shown in figure 4 with the work flow of the recognition engine on the command pattern.

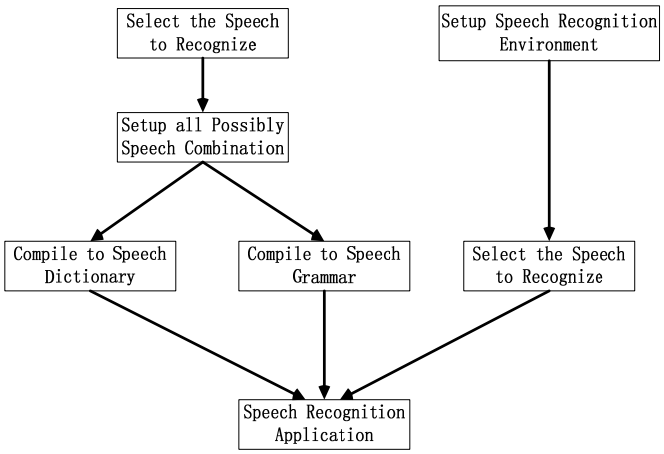

Figure 4. Working flow of the recognition engine on the command pattern.

Figure3. Speech control terminal

- The User: the operator who manipulates and commands the robot away from the sites.

- Speech recognition model: it translates the speech signal into test file for the robot's recognition after matching the user's command with the recognition program and this can be accomplished by the recognition engine and JIE DONG speech synthesize technology.

Speech recognition application: it works to receive the speech recognition, and translates it accordingly. Then the translation will be sent to a remote control end by Web Service to manipulate the robot to conduct movements in response.

IV. SPEECH FEEDBACK TTS MODEL

We have adopted the JUE DONG TTS technology in speech feedback model. jTTS is a powerful massive speech base based on real speech recording. In the process of combination, jTTS will search for the most matching unit from the massive speech base with a series of matching rules so as to make the synthesized speech as natural and smooth as possible. The user establishes the interactive file beforehand, and inputs the commands. The speech engine will search the matching file from jTTS, after the search, the engine gives out the command and sends it into the robot through internet.

As shown in figure 5, it is the work flow of a jTTS feedback. The interactive file in the system has played a very important role, functioning as a bridge between the recognition model and the application.

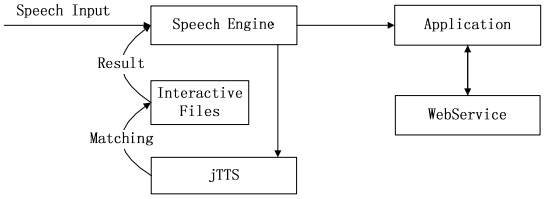

Figure 5. Working flow of a jTTS feedback.

V. REMOTE ROBOT CONTROL MODEL

Remote robot control terminal consists of two parts.

i. The remote robot: This paper has utilized a four degree of freedom industry robot, which has adopted the GT moving controller that allows it to control four movement axes and realize a multiple axis movement. Meanwhile, the controller has equipped itself the C language base and window dynamic linkage library to complete sophisticated tasks. The user then is capable of combining this language base, the data process, the interface with his control system to set up a specified application system.

ii. The robot main control system: Start the robot control program to receive and enforce the command as well as complete the real-time video collection, coding and sending tasks; the robot receives the command, and process it with application program, if the command is correct, the robot will act accordingly and will send the movement to the operator in the form of video materials to make sure the operator is well informed of the robot's work conditions.

VI. SPEECH RECOGNITION AND THE WEB SERVICES INTEGRATION

System integration solution

The Web Service in this system is based on C#, and the terminal program is compiled in VC++ 6.0. In order to develop a client program in MFC style and realize the integration with speech recognition program, it becomes a technical difficulty as to how to invoke the Web Service components in VC++ 6.0. As we know, Web Service is based on SOAP protocol and transmits data in XML format, yet the traditional C/C++ is not available for SOAP protocol and XML format. Therefore we need a third party tool to realize invoke of Web Service components. After serious research and selection, we find the gSOAP complier is able to solve this problem [6].

gSOAP complier is developed jointly by Robert van Engelen, Florida state university and Genivia. In order to set up a client to realize SOAP, we need Microsoft Soap SDK or gSOAP complier. Yet if we choose the Microsoft Soap SDK tool, the deployment of the client could be relatively complicated that some components registration is required, like the SOAP and XML.

gSOAP complier is typically used in open source project. By using gSOAP, the Client and Server programming task could be easily done in C/C++, and the programmer should not know much about xml and SOAP protocol. So the user of gSOAP could concentrate themselves on the soft programming

of the Web Service client and server's, no need to badger with other specific techs. The gSOAP is a cross-platform tool for the developing of Web Service servers and clients, coded in C/C++.

By using gSOAP complier, a user Web Service function base can be defined. The user then is able to invoke relevant functions to receive and transmit data from Web Service. This is how the author understands its working principle: the user requests remote service in the XML file format and the result is backed in XML format [6, 8].

gSOAP is better at remote data storage than DCOM and Midas as the 80 port is available in gSOAP and the deployment is rather convenient.

The process of system integration

In order to integrate the gSOAP with VC ++ 6.0 and invoke the Web Service, we need to start the following steps:

(1) Create a gSOAP Service in Visual Studio 2003 and a function Web Service. Mark the output functions with Web Method.

(2) Create a client. First, click the service and store WSDL as my service wsdl file. Then with wsdl to code, create the client code. Target the wsdl file, and click so as to generate the codes.

(3) Start a project gsoapclient, and add the ServiceSoap.nsmap，soapC.cpp，soapClient.cpp，soapH.h，soapStub.h，and stdsoap2.h，stdsoap2.cpp in the project gSOAP root directory into the project.

As shown in figure 6 with the client code flow.

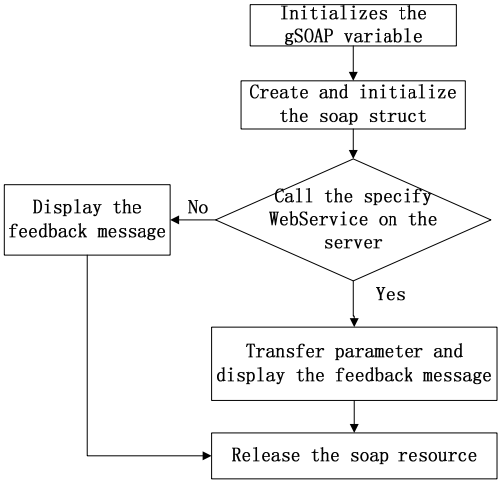

Figure 6. Logic judgement flow in client side .

Start a gsoap parameter, input and output the parameter and initialize it.

```
_ns1__shaftmove shaftmove;
_ns1__shaftmoveResponse shaftmoveResponse;
struct soap soap;
soap_init(&soap);
shaftmove.cont=cont;
shaftmove.para=para;
```

Call the soap_call___ns1__shaftmove () to invoke the Web Service on the robot control end. If the invoke is successful, it will transmit the data and display return signal. If not only the return signal is displayed. Then use the soap_cleanup(); function to release the SOAP resource.

VII. EXPERIMENTAL STUDIES

A. Experimental Design and Procedure

In order to evaluate our approach to estimate the key performance of an intelligent robot on recognition rate, we carried out an experiment in which the intelligent robot was controlled via human speech. And in the experiments, the speech data are sampled at 8 KHz and 16 bits quantization. The frame length and window shift are 23.2ms and 11.6ms, respectively. In spectra processing, after MMSE speech enhancement and spectrum smoothing, 24 triangle Mel-scaled filters are applied to combine the frequency components in each bank, and the outputs are compressed by logarithmic function. Then the Discrete cosine transform (DCT) decor relation is performed on the log-spectrum. The final acoustic feature of each frame is a 12, 24, 36 dimensional vector consisting of MFCC and their first and second order derivatives.

Seven accented speech feature databases are developed in the representative command. The 7 commands include: Shaft one turn left, Shaft one turn right, Shaft two turn left, Shaft two turn right, Shaft three turn left, Shaft three turn right, Stop, which are carefully selected according to the coverage of accents as well as economic situation. The speakers are native and the speech data are recorded in quiet environments. Each database includes 6000 utterances in training set and 100 in testing set.

The recognition algorithm can be summarized by the following steps.

Step 1: Unknown speakers' speech is recorded first.

Step 2: The starting and endpoint is detected and speech should go through the filtering process.

Step 3: Speech features are extracted from the speech signal which is used to create the testing Vector (acoustic vector) for that utterances.

Step 4: The testing vector is then fed into the vector quantizer.

Step 5: The predefined knowledge is used by the vector quantization to calculate the spectral distortion (distance) for each utterance and smallest distance value is selected.

Step 6: The smallest distance value is compared with a threshold value and a decision of whether the unknown speaker to be recognized or not is made.

B. Result analysis

The results of the experimentation are shown in Figure 7, 8,9,10 and Table 1.

By using Matlab, we analyze a single speech which from the speech features databases and get many kinds of speech signal

graphs in the form of time domains, frequency domains and point detect analyze as it shown in the following figure 7, 8, 9 mentioned in Section III.B.

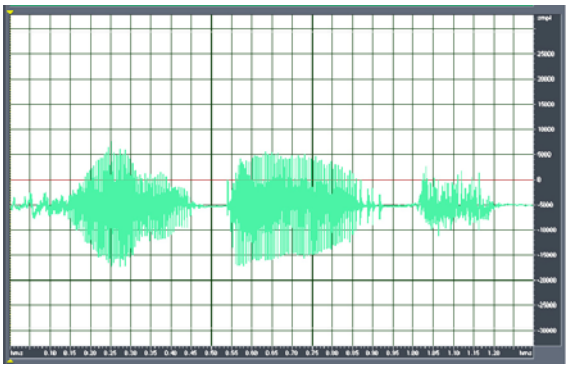

Figure 7. Speech signal in time domains.

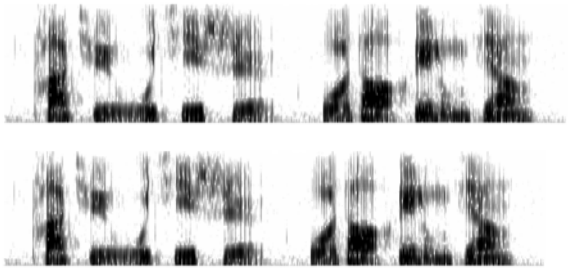

Figure8. Speech signal in frequency domains.

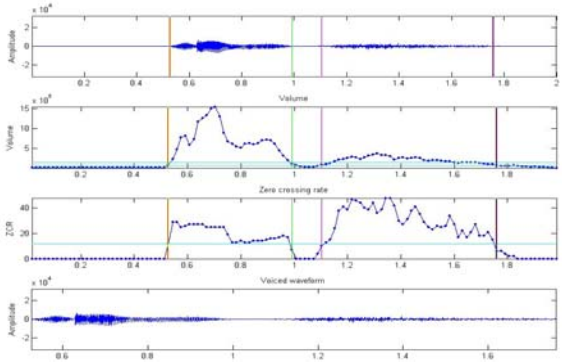

Figure 9. Point detect analyze by using Matlab

For experimental parameters, we have combined three different number of MFCC feature dimensions, as shown next.

Table 1 shows the recognize result in clean accented recognition experiments. From the results we can find that with the increase of MFCC Dimension will extend the vector quantizer time and also with the increase of speech signal frame will cause Outside Test decline greatly.

In this work, the utterances of several speakers are taken and each sample data is taken to train the vector quantizer and then all the utterances are used for recognition or testing. The input of the vector quantizer is obtained by the frequency analysis for the given input utterances. The detail of the vector quantizer is specified by representing the input in the form of Matrix. We've taken about 100 data for which the result is shown in Figure 10.

Table1.Recognize result of speech feartue data

MFCC Dimension	Frame Size	Frame overlap	Inside Test	Outside Test	VQ Time	HMM Time
12	256	171	100%	75%	0.84	0.58
12	512	171	95%	65%	0.27	0.24
12	512	85	100%	75%	0.35	0.20
12	512	171	100%	70%	0.34	0.26
24	512	85	100%	70%	0.17	0.21
24	512	171	95%	75%	0.55	0.22
24	256	171	100%	70%	0.89	0.47
36	512	171	100%	65%	2.22	0.24

Figure 10 show speech recognition rates for this subsystem, respectively. The horizontal axis indicates speech content per testing, and the vertical one indicates some key rates. By speaker saying some speeches mentioned in the speech content, we can get the speech correct recognition rates, false Inclusion rates and Rejection rates. For example, When saying "Shaft one turn right" and we can get the correct recognition rates is 83%, false inclusion rates is 6% and rejection rates 11%.

The speech recognition subsystem, we can get that the average correct recognition rate is 81.7%, the average false inclusion rate is 8.6% and the average rejection rate is 9.7%, respectively.

VIII. CONCLUSION

This paper has applied a latest speech recognition and TTS technology to realize a secondary development and develop the program available for the robot's recognition. Meanwhile, it integrates with Web Service technology to realize the remote speech control over a robot. Compared with the study of Luo, Zhizeng and Zhao, Jingbing, this robot system has a good network to use, easy remote control.

As evidence concluded from the experiment, the robot is able to reach a rate of 80% accuracy in speech recognition. As to remote control, with Web Service technology the robots' hearing can be extended profoundly, which enables in some extent the user work in a setting far away from the sites to ensure improved speech recognition accuracy. The Web Service complexes the computer-based control system, but it gives the system the distributed deployment, better flexibility and easier extension and maintenance.

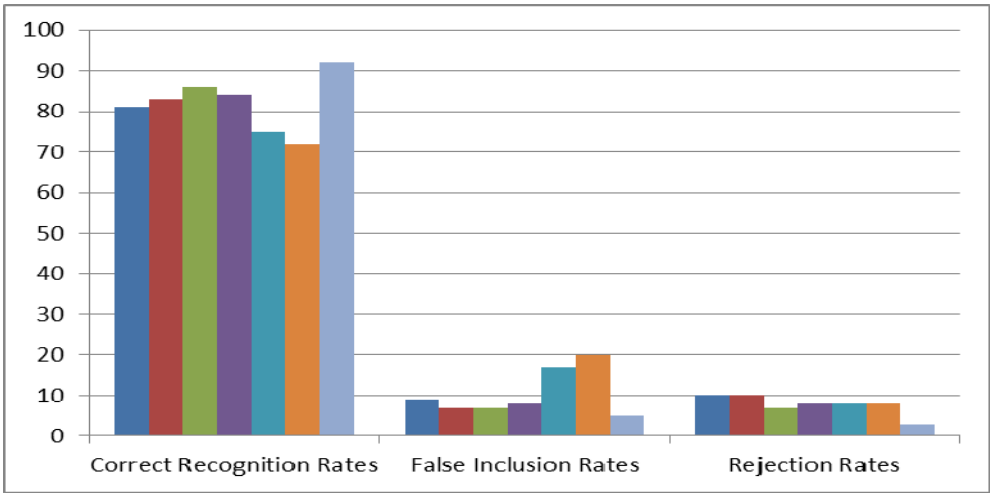

Figure 10. Recognition Rates

REFERENCES

[1] LAWRENCE R.RABINER, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” proceedings of the IEEE,vol. 77,no. 2,pp. 257–286, February 1989.

[2] L.E.Baum and T.Petrie, “Statistical inference for probabilistic functions of finite state Markov chains,” Ann.Math.Stat,vol. 37,pp. 1554–1563, April 1966.

[3] Yenan Qu, Evan F. Bollig and Gordon Erlebacher, “KWATT: a toolkit for automatic web service generation,” Visual Geosciences,vol.13, no.1, pp. 59–69, July 2008.

[4] Bojan Kotnik, Damjan Vlaj, and Bogomir Horvat, “Efficient Noise Robust Feature Extraction Algorithms for Distributed Speech Recognition (DSR) Systems,”International Journal of Speech Technology,vol. 6, no. 3,pp. 205–219, July 2003.

[5] Valérie Issarny, Daniele Sacchetti, Ferda Tartanoglu, Franoise Sailhan, Rafik Chibout, Nicole Levy,and Angel Talamona, “Developing Ambient Intelligence Systems:A Solution based on Web Services,”Automated Software Engineering,vol. 12,no. 1,pp. 101–137, 2004.

[6] G.McLachuo,Mixture Models,New York:Marcel Dokker,1998.

[7] Krishnamurthy,A.K. and D.G. Ghilders, “Two Channel Speech Analysis,” IEEE Trans. On Acoustics,Speech and Signal Processing ,vol. 34,pp. 730-743, 1986.

[8] Aphrodite Tsalgatidou and Thomi Pilioura, “An Overview of Standards and Related Technology in Web Services,” Distributed and Parallel Database,vol. 12,no. 3,pp. 135-162,Settember 2004.